Modelling Large Parallel Corpora: The Zurich Parallel Corpus Collection

Graën, Johannes ; Kew, Tannon ; Shaitarova, Anastassia ; Volk, Martin

Abstract: Text corpora come in many different shapes and sizes and carry heterogeneous annotations, depending on their purpose and design. The true benefit of corpora is rooted in their annotation and the method by which this data is encoded is an important factor in their interoperability. We have accumulated a large collection of multilingual and parallel corpora and encoded it in a unified format which is compatible with a broad range of NLP tools and corpus linguistic applications. In this paper, we present our corpus collection and describe a data model and the extensions to the popular CoNLL-U format that enable us to encode it.

DOI: https://doi.org/10.14618/ids-pub-9020

Modelling Large Parallel Corpora: The Zurich Parallel Corpus Collection

Johannes Graën[1,2], **Tannon Kew**[3], **Anastassia Shaitarova**[3], **Martin Volk**[3]

[1]Department of Swedish, University of Gothenburg
[2]Department of Translation and Language Sciences, Pompeu Fabra University
[3]Institute of Computational Linguistics, University of Zurich

Abstract

Text corpora come in many different shapes and sizes and carry heterogeneous annotations, depending on their purpose and design. The true benefit of corpora is rooted in their annotation and the method by which this data is encoded is an important factor in their interoperability. We have accumulated a large collection of multilingual and parallel corpora and encoded it in a unified format which is compatible with a broad range of NLP tools and corpus linguistic applications. In this paper, we present our corpus collection and describe a data model and the extensions to the popular CoNLL-U format that enable us to encode it.

1 Introduction

The benefit of digital corpora is rooted in their annotation. In the history of corpus linguistics, several file formats have been employed to store and distribute digital corpora. Today, we see mainly two types of corpus formats that have prevailed: a tabular one, where each line represents a token and columns contain their attributes, and a hierarchical one, where tokens are represented as leaves of a tree.

Over the years, the Institute of Computational Linguistics in Zurich has accumulated a number of large parallel corpora in different languages that span various domains and genres, have multiple layers of annotation and carry rather heterogeneous metadata. So far, corpus data has generally been stored in XML files following an ad-hoc format that has never been fully standardised but adjusted to accommodate specific characteristics and annotation. In order to standardise our corpora and to make our data directly compatible with modern Natural Language Processing (NLP) tools, we extend the CoNLL-U format (Nivre et al. 2016). Since our corpora are parallel, or have large multiparallel parts, special attention is given to the representation of alignment information. Other types of annotation, including named entities and code switching are also accounted for.

This paper first describes the theoretical relational data model that we infer from over 10 years of work on the curation of corpora, the challenges faced and our considerations regarding compatibility and extensibility (Section 2). Then we propose an extended CoNLL-U format for storing parallel corpora with multiple layers of optional annotation (Section 3). This format facilitates the aggregation of data from different corpora while being directly compatible with relational databases, allowing for complex yet efficient queries. Lastly, we present our parallel corpus collection (Section 4), which is now made available in this standardised format.

2 Data Model

First we take a high-level view of our data and create a model which considers a compositional hierarchy of three entity types: tokens, sentences and texts. The token is typically the smallest unit in text corpora (but cf. Chiarcos et al. 2012), as such, annotation is predominantly performed on tokens on a sentence-by-sentence level.[1] In our corpora, sequences of tokens form sentences, although this may not be the case for all types of corpora (e.g. Bible verses (Christodouloupoulos and Steedman 2015) or subtitles (Lison and Tiedemann 2016) which may model verses or lines). Sentences often form paragraphs, which, in turn, form coherent texts.[2] While paragraphs typically subdivide texts into smaller thematic blocks, the concept of what constitutes a paragraph is somewhat arbitrary

[1]Exceptions are methods like coreference resolution or argument detection which require annotation across a sequence of sentences.

[2]We use 'text' to refer to a cohesive and coherent body of text within a corpus that could constitute a document, article or speaker turns in parliamentary debates.

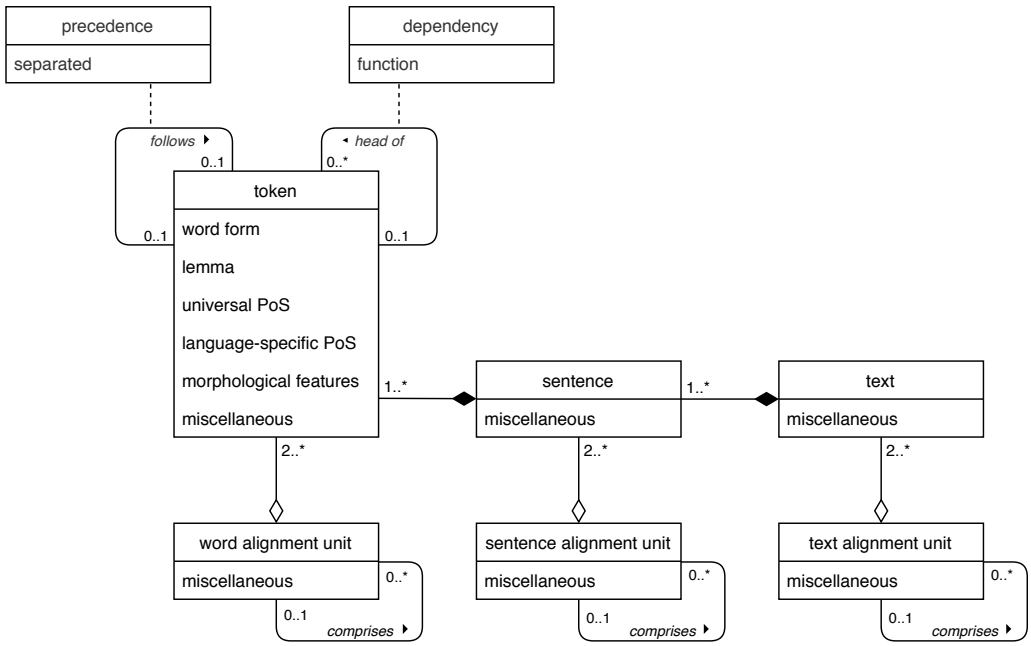


Figure 1: UML class diagram of a parallel corpus with potential hierarchical alignment on different levels.

and is often not consistently handled in different languages. Thus, we refrain from regarding paragraphs as an entity in our model, instead focusing on the hierarchy between tokens, sentences and texts (see Figure 1).

As most annotation in our corpora is centred around tokens, we model the token entity with common attributes such as surface form, lemma, part-of-speech tag and morphological features. Dependency grammar structures are represented through a recursive relationship between two tokens and defined by an attribute corresponding to the syntactic function. Here, an optional one-to-many cardinality describes dependency annotation suitable for tree structures. Graph structures can be expressed in a similar way if the source cardinality is loosened to allow for the representation of multiple heads for each token. The sequential order of tokens in a sentence is modelled as a precedence relation between two adjacent tokens. An attribute of this relation specifies whether tokens are separated by white space in the original surface form of a sentence, allowing for accurate reconstruction.

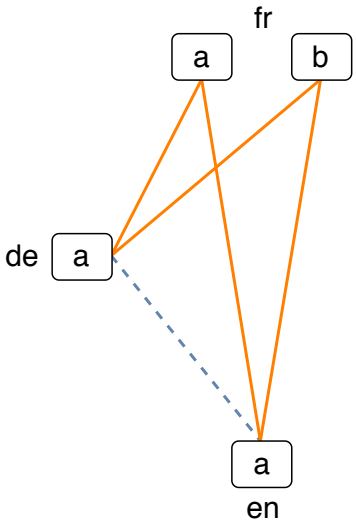
A 'miscellaneous' attribute at each level of the hierarchy allows for any relevant, unstructured information to be stored. For instance, to model both inter-sentential and intra-sentential code-switching (see Volk and Clematide 2014), we use this field to mark a token when its language deviates from that of its sentence and, similarly, for a sentence when its language differs from that of its text. While sentence and text entity types generally demand far fewer levels of annotation than tokens, the miscellaneous attribute permits arbitrary metadata, for example, formatting and layout information at the sentence level or speaker attribution at the text level.

2.1 Modelling Alignment

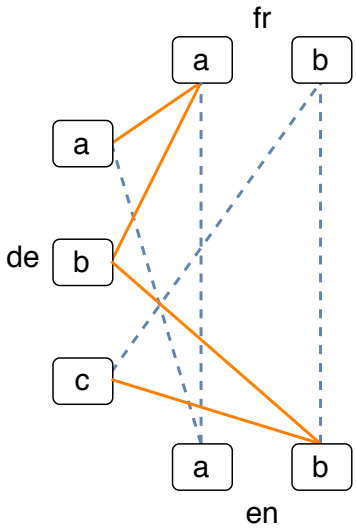
Alignment is modelled on token, sentence and text level as the affiliation of an entity to an alignment unit. This allows multilingual hierarchical alignment (Graën 2018, Sections 4.3 and 4.5) to be represented the same way as regular bilingual alignment. In most of our corpora, alignments are primarily bilingual.[3] In order to obtain multilingual alignments, we aggregate all corresponding bilingual alignments.[4] However, as illustrated in Figure 2, this approach does not always yield coherent and meaningful alignments across all languages. Figure 2a shows the ideal scenario, where the combination of one-to-one and one-to-many alignments is coherent, while in Figure 2b the combination results in an incoherent multilingual alignment. Nevertheless, modelling alignments in this way makes it possible to extract a subset of the available languages from any alignment unit.

[3] Except for the Sparcling corpus, which contains multilingual text and sentence alignments (Graën 2018).

[4] An alternative approach to representing multilingual alignment is to rely on a 'pivot' language (see Steinberger et al. 2014; Zeroual and Lakhouaja 2018).



(a) Coherence between one-to-one and one-to-many alignments.



(b) Incoherence between one-to-one and one-to-many alignments

Figure 2: Multiparallel alignment based on combining pairwise alignments with one-to-one relations (blue dashed edges) and one-to-many relations (orange solid edges).

3 Encoding Parallel Corpora

3.1 A smorgasbord of corpus formats

One of the most widely adopted approaches to encoding text corpora is XML (eXtensible Markup Language), which allows for a hierarchical representation using a tree structure. Such a representation is valuable for the storage of language data as it facilitates the clear separation of structural information from text content, provides a descriptive markup of the encoded text, and can easily be validated for consistency with an appropriate document type definition (DTD) or XML schema. For this reason, groups such as the Text Encoding Initiative[5] (TEI) have establish a standardised specification for the encoding of text corpora in XML (see also Dipper 2005; Hana and Štěpánek 2012; Gompel and Reynaert 2013).

A second approach is the tabular format that has quickly gained popularity and become the de facto standard in the NLP community (Buchholz and Marsi 2006; Chiarcos and Schenk 2018). The CoNLL-U format (Nivre et al. 2016) defines a standardised method of encoding text corpora for Universal Dependency (UD) Treebanks. It is based on a simple one-word-per-line (OWPL) format in which annotation layers are stored in ten distinct columns and are thus defined by their position, rather than markup tags.[6] This lightweight format is reminiscent of that used by the IMS Open Corpus Workbench (CWB) (Evert and the CWB Development Team 2010), which is able to blend both structural XML tags, albeit without being valid XML, and a tabular representation of a token's attributes in order to encode only the necessary linguistic information for a given task. Additionally, multiple extensions have been proposed to the basic CoNLL-U format, for example, for the annotation of multiword expressions (Savary et al. 2017) and morphological analysis (More et al. 2018). A more recent dialect of the CoNLL family is the CoNLL-U Plus format, which defines a modified CoNLL-U file that can contain any number of columns to flexibly encode any additional linguistic annotations while still maintaining a valid CoNLL format.

3.2 One format to rule them all

Despite the large number of corpus formats, there is little support for the representation of alignments. We decide to encode our corpora in what is essentially a CoNLL-U format and extend it with optional layers of stand-off annotation to accommodate the data model described in Section 2. Figure 3 depicts an excerpt from a corpus with multilingual alignments.

[5] https://tei-c.org/

[6] Numerous versions of the CoNLL format exist due to its application in multiple shared tasks held by the Conference on Computational Natural Language Learning (CoNLL) since 2006. CoNLL-U is an extension of CoNLL-X/CoNLL-ST which were themselves extensions of Joakim Nivre's MaltTAB format (Buchholz and Marsi 2006).

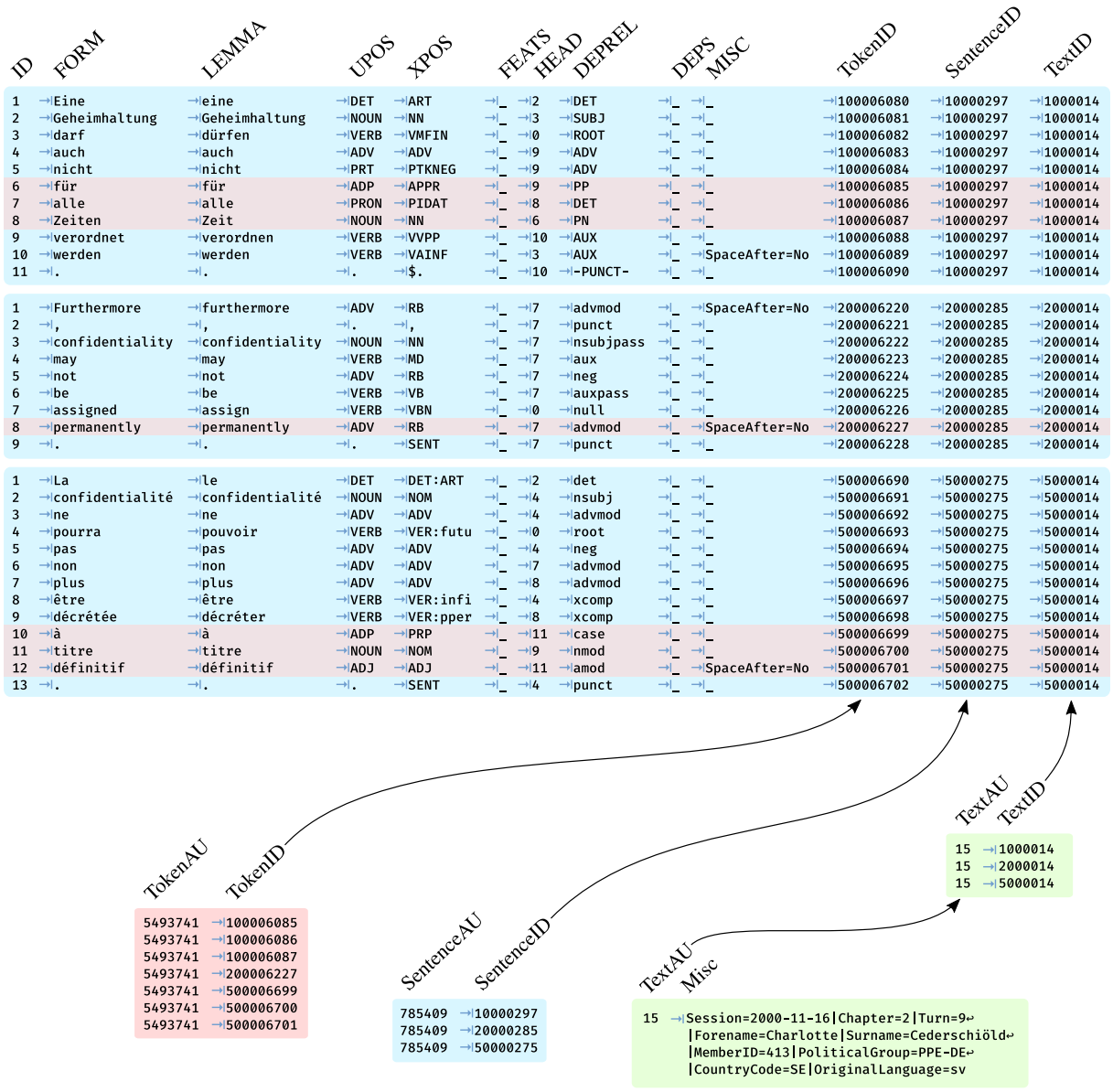


Figure 3: An excerpt of our extended CoNLL-U format for a parallel corpus with multilingual alignments. Snippets of stand-off files show token, sentence and text alignments. As depicted here, language-independent meta information can also be attached to alignment units.

Adopting CoNLL-U as a basis for our corpora brings a number of advantages: i) it ensures direct compatibility with numerous NLP tools[7], including state-of-the-art taggers and parsers, thereby making it easy to re-annotate our corpora as systems improve; ii) it guarantees that our corpora are directly compatible with relational database systems allowing for complex corpus queries; iii) it is human-readable and facilitates the extraction of language and task-specific data using simple command-line tools (e.g. `grep`, `sed`, `awk`); and iv) it provides a standardised base format for our large multilingual corpus collection, allowing for cross compatibility between corpora and serving as a good starting point for conversions into other transfer formats (e.g. TEI).

Naturally, there are some obvious shortcomings related to opting for a simplified tabular format to encode text corpora, some of which are discussed by Straňák and Štěpánek (2010) in their critique of the early CoNLL format. For example: i) multiple levels of sparse annotation can quickly lead to unwieldy tables; ii) corpus validation is made more difficult due to the lack of a DTD or schema for ensuring consistency; and iii) the inclusion of metadata and layout information, such

[7] https://universaldependencies.org/tools.html

as the placement of HTML tags, page breaks or graphics, which may be relevant for some analyses or veracity evaluation, is made difficult and cumbersome when moving away from XML markup.

3.3 Our Format

We split our corpora into language-specific subsections. For each section, tokens are furnished with ubiquitous attributes, pertaining to those specified by CoNLL-U in a main tabular token file.[8] These attributes include a sentence-positional identifier (word index), surface form, lemma, part-of-speech tags, morphological features, information for dependency relations and a miscellaneous attribute for additional token-level annotation. Unspecified or empty values are represented by an underscore ('_'). In the miscellaneous column, a list of attribute-value pairs is used to hold corpus-specific annotations at the token level (in the form of attribute=value, separated by pipe ('|') characters). In addition to the 10 columns defined by CoNLL-U, we include three enumerated identifier (ID) values. These IDs comprise one (primary) key, which uniquely identifies each token in a corpus, and two (foreign) keys, which reference the token's corresponding sentence and document.[9] All IDs are expected to increase linearly throughout the file, which facilitates processing.

Sentence-level and text-level annotations are then stored separately with relevant metadata based on their enumerated IDs. For consistency, we follow the same approach as in the token file and include a miscellaneous attribute for sentences and texts with a list of attribute-value pairs. Finally, we specify additional stand-off annotation files in order to accommodate non-ubiquitous annotation such as named entities and multilingual alignment. As such, stand-off files are only required when those annotations are present.

4 The Zurich Parallel Corpus Collection

Having brought our parallel corpus collection into a consistent and standardised format, as described in Section 3, we make these resources publicly available. This corpus collection provides a rich source of multilingual and multiparallel language data in a variety of domains and genres. A brief overview of the collection is given in Table 1.

At the heart of our collection lies the heritage corpus of alpine texts, **Text+Berg**[10] (Volk et al. 2010; Göhring and Volk 2011). This corpus consists of 150 years of digitised material from the Swiss Alpine Club yearbooks, which were published primarily in German and French, with some years containing texts in Italian, Romansh, English and also Swiss German.[11] Approximately 15% of the corpus comprises a German-French parallel subsection of roughly 4.5 million tokens per language. Over 10 years in development, Text+Berg has inspired numerous innovative approaches in corpus annotation, such as crowd-sourced correction of OCR errors (Clematide, Furrer et al. 2016), named entity recognition and linking (Ebling et al. 2011), code-switching (Volk and Clematide 2014), and special handling of elliptical compound nouns and separable prefix verbs in German (Volk, Clematide et al. 2016).

The **Credit Suisse Bulletin** corpus (CS Bulletin)[12] (Volk, Amrhein et al. 2016) is based on the world's oldest banking magazine published by Credit Suisse. This magazine has been in print since 1895 in both German and French, with translations also produced in English, Italian and Spanish at certain periods. There are more than 20 million tokens in the German and the French part, while the English and Italian sections contain about 10 million tokens per language. The Credit Suisse Bulletin corpus provides parallel data from magazine articles in the domains of economics, culture and sport, proving to be useful material for historic, sociological and linguistic research (Schneider et al. 2018).

The **Swiss Legislation Corpus** (SLC) (Höfler and Sugisaki 2014) is a German-French parallel corpus comprised of the entire classified collection of contemporary legislative writing of the Swiss Confederation. Its companion, the **Rumantsch Grischun corpus**[13] (Weibel 2014), consists of legal texts and press releases from the State Chancellery of the Swiss canton of Graubünden. This corpus provides unique parallel data for German and the low-resource language Romansh. As such, it is a valuable resource for Romansh language

[8] A header comment line beginning with '#' defines the columns and relevant namespaces, ensuring that it conforms with CoNLL-U Plus.

[9] Primary and foreign keys are terms borrowed from database design.

[10] http://textberg.ch/

[11] Although Swiss German has no official written standard, it is often written by native speakers in non-formal situations.

[12] https://pub.cl.uzh.ch/projects/b4c/en/

[13] 'Rumantsch' is an alternative spelling of 'Romansh'.

	languages	tokens	years	alignment
Text+Berg	de, fr, it, rm, gsw, en	52.6m	150	sentence
CS Bulletin	de, en, es, fr, it	61.6m	120	sentence
Sparcling	de, en, es, fr, it + 11	454.7m	15	token
SLC	de, fr	11.4m	—	token
Rumantsch Grischun	de, rm	0.9m	—	token
Medi-Notice	de, fr, it	58.9m	—	sentence
Horizons	de, en, fr	2.9m	14	text

Table 1: List of corpora together with their most relevant characteristics.

learners and a solid base for computational linguistic research.

The largest multiparallel corpus in our collection is the **Sparcling** corpus, originally referred to as FEP9 (Graën 2018). Sparcling is a richly annotated development of the CoStEP corpus (Graën et al. 2014), which itself is a cleaned and normalised version of the Europarl corpus (Koehn 2005). Token counts for each language vary, ranging from 7.5 to 47 million across the 16 languages, with annotation and alignment on all levels. Thus, it provides a rich resource for comparative language studies (Callegaro 2017), language learning applications (Schneider and Graën 2018) and the development of multilingual NLP methods (Heierli 2018). It has also been used in the implementation of a query and exploration system for multiparallel corpora (Clematide, Graën et al. 2016; Graën et al. 2017).

The **Medi-Notice** corpus (Fritz 2016) comprises texts from information leaflets for pharmaceutical products that are made publicly available by the Swiss Agency for Therapeutic Products. Each product usually has two separate leaflets: one is geared towards medical professionals, while the other is written for the general public. According to Swiss law, patient leaflets must be written in German, French and Italian, whereas the information for healthcare professionals is required only in German and French. Thus, the Medi-Notice corpus contains German and French parallel texts in the professional subsection, while the patient subsection is trilingual.

Lastly, the **Horizons** corpus[14] is a multiparallel corpus constructed from the magazine of the same name, published by the Swiss National Science Foundation.[15] This corpus also offers unique parallel texts in the domain of popular science in and around Switzerland in German, French and English.

5 Conclusions and Future Development

Through the development of the corpora mentioned above and the challenges involved in handling large multiparallel corpora, we have deduced a data model which allows us to represent the diversity of annotations in our corpora effectively. We have extended the CoNLL-U format to encode our corpora, which ensures compatibility with modern NLP applications and corpus linguistic tools, facilitates the extraction and the exploitation of linguistic data, and allows extensibility through various layers of stand-off annotation. Additionally, we have made our corpora available in this format, totalling approximately 640 million tokens across 18 languages. We hope that this will enable a more effective and efficient application of multiparallel corpora in a variety of linguistic research projects. At present, we are working on tools to handle corpora in our tabular format. This includes validation of the corpus files, extraction of task-specific subsections and conversion pipelines into other formats such as TEI. Further information and the corpus files are available at `https://pub.cl.uzh.ch/purl/PaCoCo`.

6 Acknowledgements

We would like to acknowledge the many contributors who have helped to develop the Zurich Parallel Corpus Collection described in this paper. Their extensive and valuable efforts over many years have made this current work possible. We would also like to thank the anonymous reviewers for their helpful comments and suggestions.

[14] The Horizons corpus has not yet been officially published and development is still underway, but it is being made available in its current form as part of this release.

[15] `https://www.horizons-mag.ch/`

References

Buchholz, Sabine and Erwin Marsi (2006). 'CoNLL-X Shared Task on Multilingual Dependency Parsing'. In: *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*. New York, pp. 149–164.

Callegaro, Elena (2017). 'Parallel Corpora for the Investigation of (Variable) Article Use in English: A Construction Grammar Approach'. PhD thesis. University of Zurich.

Chiarcos, Christian, Julia Ritz and Manfred Stede (2012). 'By all these lovely tokens… Merging Conflicting Tokenizations'. In: *Proceedings of the Linguistic Annotation Workshop (LAW)*, pp. 53–74.

Chiarcos, Christian and Niko Schenk (2018). 'The ACoLi CoNLL Libraries: Beyond Tab-Separated Values'. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*. Ed. by Nicoletta Calzolari et al., pp. 571–576.

Christodouloupoulos, Christos and Mark Steedman (2015). 'A massively parallel corpus: the Bible in 100 languages'. In: *Language Resources and Evaluation* 49.2, pp. 375–395.

Clematide, Simon, Lenz Furrer and Martin Volk (2016). 'Crowdsourcing an OCR Gold Standard for a German and French Heritage Corpus'. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portoroz, Slovenia, pp. 975–982.

Clematide, Simon, Johannes Graën and Martin Volk (2016). 'Multilingwis – A Multilingual Search Tool for Multi-Word Units in Multiparallel Corpora'. In: *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives – Fraseologia computacional y basada en corpus: perspectivas monolingües y multilingües*. Ed. by Gloria Corpas Pastor. Geneva: Tradulex, pp. 447–455.

Dipper, Stefanie (2005). 'XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation'. In: *Proceedings of Berliner XML Tage*, pp. 39–50.

Ebling, Sarah, Rico Sennrich, David Klaper and Martin Volk (2011). 'Digging for Names in the Mountains: Combined Person Name Recognition and Reference Resolution for German Alpine Texts'. In: *5th Language & Technology Conference (LTC)*, pp. 189–200.

Evert, Stefan and the CWB Development Team (2010). *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial*.

Fritz, Andrea (2016). 'Erstellung eines parallelen Arzneimittelinformations-Korpus (Deutsch-Französisch) und Optimierung von dafür einsetzbaren Part-of-Speech-Taggern'. MA thesis. University of Zurich.

Göhring, Anne and Martin Volk (2011). 'The Text+Berg Corpus – An Alpine French-German Parallel Resource'. In: *Traitement Automatique des Langues Naturelles*, pp. 63–68.

Gompel, Maarten van and Martin Reynaert (2013). 'FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study'. In: *Computational Linguistics in the Netherlands* 3, pp. 63–81.

Graën, Johannes (2018). 'Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning'. PhD thesis. University of Zurich.

Graën, Johannes, Dolores Batinic and Martin Volk (2014). 'Cleaning the Europarl Corpus for Linguistic Applications'. In: *Proceedings of the 12th Conference on Natural Language Processing (KONVENS)*, pp. 222–227.

Graën, Johannes, Dominique Sandoz and Martin Volk (2017). 'Multilingwis2 – Explore Your Parallel Corpus'. In: *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*. Linköping Electronic Conference Proceedings 131, pp. 247–250.

Hana, Jirka and Jan Štěpánek (2012). 'Prague Markup Language Framework'. In: *Proceedings of the 6th Linguistic Annotation Workshop (LAW)*. Jeju, Republic of Korea, pp. 12–21.

Heierli, Jasmin (2018). 'Lemma Disambigation in Multilingual Parallel Corpora'. MA thesis. University of Zurich.

Höfler, Stefan and Kyoko Sugisaki (2014). 'Constructing and Exploiting an Automatically Annotated Resource of Legislative Texts'. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. Ed. by Nicoletta Calzolari et al., pp. 175–180.

Koehn, Philipp (2005). 'Europarl: A parallel corpus for statistical machine translation'. In: *Proceedings of the 10th Machine Translation Sum-*

mit. Vol. 5. Asia-Pacific Association for Machine Translation (AAMT), pp. 79–86.

Lison, Pierre and Jörg Tiedemann (2016). 'OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles'. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Ed. by Nicoletta Calzolari et al., pp. 923–929.

More Amirand Çetinoğlu, Özlem, Çağrı Çöltekin, Nizar Habash, Benoît Sagot, Djamé Seddah, Dima Taji and Reut Tsarfaty (2018). 'CoNLL-UL: Universal Morphological Lattices for Universal Dependency Parsing'. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*. Ed. by Nicoletta Calzolari et al., pp. 3847–3853.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty and Daniel Zeman (2016). 'Universal Dependencies v1: A Multilingual Treebank Collection'. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Ed. by Nicoletta Calzolari et al., pp. 1659–1666.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati and Veronika Vincze (2017). 'The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions'. In: *Proceedings of the 13th Workshop on Multiword Expressions*. Valencia, Spain, pp. 31–47.

Schneider, Gerold and Johannes Graën (2018). 'NLP Corpus Observatory – Looking for Constellations in Parallel Corpora to Improve Learners' Collocational Skills'. In: *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*, pp. 69–78.

Schneider, Gerold, Anastassia Shaitarova and Martin Volk (2018). 'Credit Suisse Bulletin Corpus: The world's Oldest Banking Magazine as a Treasure Trove of Applications for Digital Humanities'. Poster at Workshop DARIAH-CH. University of Neuchâtel.

Steinberger, Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski and Signe Gilbro (2014). 'An overview of the European Union's highly multilingual parallel corpora'. In: *Language Resources and Evaluation* 48.4, pp. 679–707.

Straňák, Pavel and Jan Štěpánek (2010). 'Representing Layered and Structured Data in the CoNLL-ST Format'. In: *Proceedings of the 2nd International Conference on Global Interoperability for Language Resources (ICGL)*, pp. 143–152.

Volk, Martin, Chantal Amrhein, Noëmi Aepli, Mathias Müller and Phillip Ströbel (2016). 'Building a Parallel Corpus on the World's Oldest Banking Magazine'. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, pp. 288–296.

Volk, Martin, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer and Beni Ruef (2010). 'Challenges in Building a Multilingual Alpine Heritage Corpus'. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*. Ed. by Nicoletta Calzolari et al.

Volk, Martin and Simon Clematide (2014). 'Detecting Code-Switching in a Multilingual Alpine Heritage Corpus'. In: *Proceedings of the 1st Workshop on Computational Approaches to Code Switching*. Doha, Qatar, pp. 24–33.

Volk, Martin, Simon Clematide, Johannes Graën and Phillip Ströbel (2016). 'Bi-particle Adverbs, PoS-Tagging and the Recognition of German Separable Prefix Verbs'. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*, pp. 297–305.

Weibel, Manuela (2014). 'Aufbau paralleler Korpora und Implementierung eines wortalignierten Suchsysteme für Deutsch – Rumantsch Grischun'. MA thesis. University of Zurich.

Zeroual, Imad and Abdelhak Lakhouaja (2018). 'MulTed: A Multilingual Aligned and Tagged Parallel Corpus'. In: *Applied Computing and Informatics*.